5-00059-TS

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-282321
(43)Date of publication of application : 31.10.1997

(51)Int.Cl. G06F 17/28
G06F 17/27
G10L 3/00

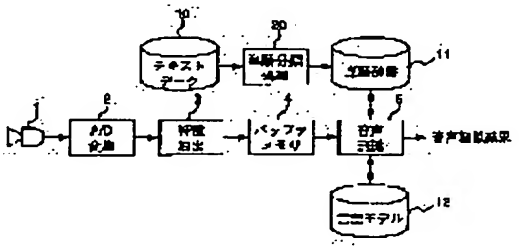
(21)Application number : 08-198950
(22)Date of filing : 29.07.1996

(71)Applicant : ATR ONSEI HONYAKU TSUSHIN KENKYUSHO:KK
(72)Inventor : SHIODA AKIRA
IIDA HITOSHI

(30)Priority
Priority number : 08 27809 Priority date : 15.02.1996 Priority country : JP

(54) WORD CLASSIFICATION PROCESSING METHOD AND DEVICE THEREFOR, AND VOICE RECOGNIZER

(57)Abstract:
PROBLEM TO BE SOLVED: To obtain a word classification result having a well-balanced hierarchical structure by classifying plural words into plural classes in the form of a binary tree having hierachized lower, intermediate and upper layers.
SOLUTION: The word classification processing part 20 classifies the words included in the text data stored in a text data memory 10 by assigning the words of comparatively low appearance frequency and the words of high rates to be adjacent to the same word in the same classes respectively. Then, the part 20 classes the word classification result into the intermediate, upper and lower layers. Then, the words are classified in order of intermediate, upper and lower layers and based on the prescribed average mutual information content, i.e., a global (overall) cost function set for all words included in the text data. The classified words are stored in a word dictionary memory 11 in the form of a word dictionary. In such word classification processing, it is possible to obtain the word classification result that has a well-balanced hierarchical structure and also is globally optimized.



LEGAL STATUS

[Date of request for examination] 29.07.1996

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]	3043625
[Date of registration]	10.03.2000
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－282321

(43)公開日　平成９年(1997)10月31日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	17/28			Ｇ０６Ｆ	15/38	Ｃ
	17/27			Ｇ１０Ｌ	3/00	５６１Ｇ
Ｇ１０Ｌ	3/00	５６１		Ｇ０６Ｆ	15/38	Ｅ

審査請求　有　　請求項の数５　ＯＬ　（全 22 頁）

(21)出願番号　特願平8－198950

(22)出願日　平成８年(1996)７月29日

(31)優先権主張番号　特願平8－27809

(32)優先日　平８(1996)２月15日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　593118597
株式会社エイ・ティ・アール音声翻訳通信研究所
京都府相楽郡精華町大字乾谷小字三平谷５番地

(72)発明者　潮田　明
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール音声翻訳通信研究所内

(72)発明者　飯田　仁
京都府相楽郡精華町大字乾谷小字三平谷５番地　株式会社エイ・ティ・アール音声翻訳通信研究所内

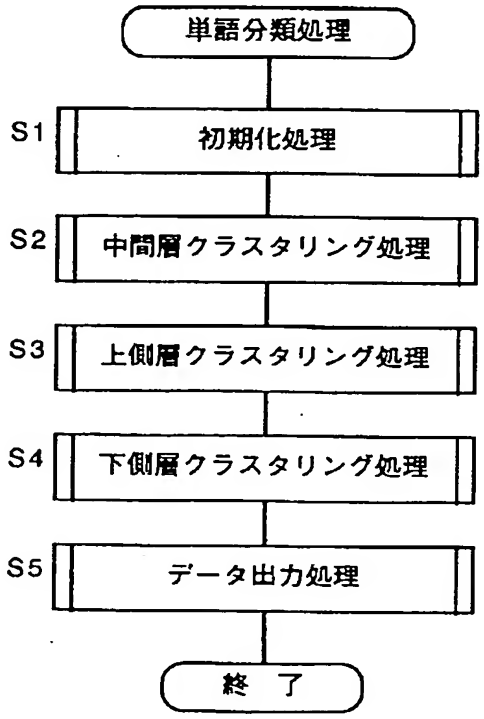
(74)代理人　弁理士　青山　葆　（外２名）

(54)【発明の名称】　単語分類処理方法、単語分類処理装置及び音声認識装置

(57)【要約】

【課題】　単語分類処理によりバランスのとれた階層構造を有しかつ全体的に最適化された単語分類結果を得ることができる単語分類処理方法、単語分類処理装置、及びその単語分類処理装置を備えた音声認識装置を提供する。

【解決手段】　複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当て、上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１のメモリに記憶し、当該クラスの単語に基づいて、所定の平均相互情報量が最大となるように、複数の単語を二分木の形式で複数ｃ個のクラスに分類して単語分類結果を表わす全体のツリー図の中間層を求め、当該中間層を中心として、上側層と、中間層の各クラス毎の複数の下側層とを求めて、全体のツリー図を求める。



【特許請求の範囲】

【請求項１】　複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当てるステップと、

上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１の記憶装置に記憶するステップと、

上記第１の記憶装置に記憶された１つのウィンドウ内のクラスの単語に基づいて、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する確率を、上記第１のクラスの単語の出現確率と第２のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数ｃ個のクラスに分類し、分類された複数ｃ個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数ｃ個のクラスとして第２の記憶装置に記憶するステップと、

上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第３の記憶装置に記憶するステップと、

上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスの各クラス毎に、上記中間層の複数ｃ個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第４の記憶装置に記憶するステップと、

上記第４の記憶装置に記憶された上記ツリー図の下側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結する一方、上記第３の記憶装置に記憶された上記ツリー図の上側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第５の記憶装置に記憶するステップとを備えたことを特徴とする単語分類処理方法。

【請求項２】　上記分類された複数ｃ個のクラスを上記第２の記憶装置に記憶するステップは、上記第１の記憶装置に記憶された１つのウィンドウよりも外側のクラスが存在し、又は上記１つのウィンドウ内のクラスがｃ個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする請求項１記載の単語分類処理方法。

【請求項３】　複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当てる第１の制御手段と、

上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１の記憶装置に記憶する第２の制御手段と、

上記第１の記憶装置に記憶された１つのウィンドウ内のクラスの単語に基づいて、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する確率を、上記第１のクラスの単語の出現確率と第２のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数ｃ個のクラスに分類し、分類された複数ｃ個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数ｃ個のクラスとして第２の記憶装置に記憶する第３の制御手段と、

上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第３の記憶装置に記憶する第４の制御手段と、

上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスの各クラス毎に、上記中間層の複数ｃ個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第４の記憶装置に記憶する第５の制御手段と、

上記第４の記憶装置に記憶された上記ツリー図の下側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結する一方、上記第３の記憶装置に記憶された上記ツリー図の上側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第５の記憶装置に記憶する第６の制御手段とを備えたことを特徴とする単語分類処理装置。

【請求項４】　上記第３の制御手段は、上記第１の記憶装置に記憶された１つのウィンドウよりも外側のクラスが存在し、又は上記１つのウィンドウ内のクラスがｃ個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする請求項３記載の単語分類処理装置。

【請求項５】　入力される発声音声の音声信号に基づいて、請求項３又は４記載の単語分類処理装置によって複数の単語が複数のクラスに分類された単語分類結果を含む単語辞書と、所定の隠れマルコフモデルとを参照して

上記発声音声を音声認識する音声認識手段を備えたことを特徴とする音声認識装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、音声認識装置、形態素解析装置、及び構文解析装置のための単語分類処理方法及び単語分類処理装置、並びに、上記単語分類処理装置を備えた音声認識装置に関する。

【０００２】

【従来の技術】単語の分類体系は、音声認識装置、形態素解析装置や構文解析装置において処理を円滑に行う上で非常に重要な知識の１つである。この単語の分類体系を構築するための１つの方法として、大量のテキストデータに基づいて単語間の相互情報量を用いた方法（以下、第１の従来例という。）が、例えば、従来技術文献「Peter Brown, et al, "Class-Based n-gram Models of Natural Language", Computational Linguistics, Vol.18, No.4, pp.467-479, １９９２年」において提案されている。この従来例の方法においては、ｎ－グラムモデルを用いて所定の相互情報量を計算して英単語の分類を行っている。

【０００３】しかしながら、第１の従来例の相互情報量による分類処理方法を用いて英単語を分類した場合、出現頻度の低い単語が不適切に分類される場合が多いという問題点があった。この問題点を解決するために、単語分類処理装置及び音声認識装置（以下、第２の従来例という。）が、本出願人により特願平７－０５６９１８号の特許出願において提案されている。

【０００４】当該第２の従来例の単語分類処理装置は、単語のｎ－グラムを利用して、同一の単語に隣接する割合の多い単語を同一のクラスに割り当てるという基準で複数の単語を複数のクラスに分類する第１の分類手段と、上記第１の分類手段によって分類された複数の単語に対して、すべての単語の出現頻度を調べ、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する頻度を、上記第１のクラスの単語の出現頻度と第２のクラスの単語の出現頻度との積に対する相対的な頻度の割合を表わす所定の相互情報量が最大となるように、上記複数の単語を二分木の形式で複数のクラスに分類する第２の分類手段とを備えたことを特徴としている。ここで、上記第２の分類手段は、好ましくは、上記第１の分類手段によって分類された複数の単語に対して、すべての単語の出現頻度を調べ、出現頻度の高い単語から順に、所定の複数Ｎ個のクラスに割り当て、Ｎ個のクラスの中で上記相互情報量が最大である２つのクラスを１つのクラスとしてまとめることにより、（Ｎ－１）個のクラスに分類し、クラスに割り当てられていない単語の中で、出現頻度が最大のものを新たにＮ番目のクラスとして割り当て、すべての単語がＮ個のクラスに割り当てられるまで、上記の処理を繰り返し、現在あるクラスから上記相互情報量が最大である２つのクラスを１つのクラスとしてまとめ、この処理を１個のクラスにまとまるまで繰り返す。これにより、単語分類をより安定に実行することができ、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができるという特徴を有する。

【０００５】

【発明が解決しようとする課題】上記第１の従来例の相互情報量による分類処理方法を用いて英単語を分類した場合、出現頻度の低い単語が不適切に分類される場合が多い。この原因としては、分離結果がバランスのとれた階層構造となっていないためであると考えられる。

【０００６】また、上記第２の従来例においては、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する頻度を、上記第１のクラスの単語の出現頻度と第２のクラスの単語の出現頻度との積に対する相対的な頻度の割合を表わす所定の相互情報量が最大となるように、上記複数の単語を二分木の形式で複数のクラスに分類しているので、上記第１のクラスの単語と上記第２のクラスの単語においては、局所的に最適化された単語分類結果を得ることができるが、全体的に最適化された単語分類結果を得ることができないという問題点があった。

【０００７】本発明の目的は以上の問題点を解決し、単語分類処理によりバランスのとれた階層構造を有しかつ全体的に最適化された単語分類結果を得ることができる単語分類処理方法、単語分類処理装置、及びその単語分類処理装置を備えた音声認識装置を提供することにある。

【０００８】

【課題を解決するための手段】本発明に係る請求項１記載の単語分類処理方法は、複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当てるステップと、上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１の記憶装置に記憶するステップと、上記第１の記憶装置に記憶された１つのウィンドウ内のクラスの単語に基づいて、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する確率を、上記第１のクラスの単語の出現確率と第２のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数ｃ個のクラスに分類し、分類された複数ｃ個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数ｃ個のクラスとして第２の記憶装置に記憶するステップと、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を

二分木の形式で１個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第３の記憶装置に記憶するステップと、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスの各クラス毎に、上記中間層の複数ｃ個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第４の記憶装置に記憶するステップと、上記第４の記憶装置に記憶された上記ツリー図の下側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結する一方、上記第３の記憶装置に記憶された上記ツリー図の上側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第５の記憶装置に記憶するステップとを備えたことを特徴とする。

【０００９】また、請求項２記載の単語分類処理方法は、請求項１記載の単語分類処理方法において、上記分類された複数ｃ個のクラスを上記第２の記憶装置に記憶するステップは、上記第１の記憶装置に記憶された１つのウィンドウよりも外側のクラスが存在し、又は上記１つのウィンドウ内のクラスがｃ個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする。

【００１０】本発明に係る請求項３記載の単語分類処理装置は、複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当てる第１の制御手段と、上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１の記憶装置に記憶する第２の制御手段と、上記第１の記憶装置に記憶された１つのウィンドウ内のクラスの単語に基づいて、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する確率を、上記第１のクラスの単語の出現確率と第２のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数ｃ個のクラスに分類し、分類された複数ｃ個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数ｃ個のクラスとして第２の記憶装置に記憶する第３の制御手段と、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第３の記憶装置に記憶する第４の制御手段と、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスの各クラス毎に、上記中間層の複数ｃ個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第４の記憶装置に記憶する第５の制御手段と、上記第４の記憶装置に記憶された上記ツリー図の下側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結する一方、上記第３の記憶装置に記憶された上記ツリー図の上側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第５の記憶装置に記憶する第６の制御手段とを備えたことを特徴とする。

【００１１】また、請求項４記載の単語分類処理装置は、請求項３記載の単語分類処理装置において、上記第３の制御手段は、上記第１の記憶装置に記憶された１つのウィンドウよりも外側のクラスが存在し、又は上記１つのウィンドウ内のクラスがｃ個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする。

【００１２】本発明に係る請求項５記載の音声認識装置は、入力される発声音声の音声信号に基づいて、請求項３又は４記載の単語分類処理装置によって複数の単語が複数のクラスに分類された単語分類結果を含む単語辞書と、所定の隠れマルコフモデルとを参照して上記発声音声を音声認識する音声認識手段を備えたことを特徴とする。

【００１３】

【発明の実施の形態】以下、図面を参照して本発明に係る実施形態について説明する。図１は、本発明に係る第１の実施形態の音声認識装置のブロック図である。この音声認識装置は、テキストデータメモリ１０内のテキストデータ内の単語について出現頻度の比較的低い単語を、同一の単語に隣接する割合の多い単語を同一のクラスに割り当てるという基準で分類した後、単語分類結果を中間層、上側層、及び下側層の３つの階層に分類し、テキストデータ内のすべての単語を対象とするグローバルな（全体的な）コスト関数である所定の平均相互情報量を用いて、中間層、上側層、及び下側層の順序で階層別に単語の分類を実行して、単語辞書メモリ１１内に単語辞書として格納する単語分類処理部２０を備えたことを特徴とする。

【００１４】＜単語分類処理方法＞まず、本発明に係る本実施形態の単語の分類（クラスタリング）方法について、第１の従来例の方法と対比させて説明する。本発明の方法は、従来技術文献に開示された第１の従来例の方

法を修正しかつ大幅に発展させて改善させた方法であって、第１の従来例の式と、本実施形態の式との相違について説明し、次いで、単語の分類処理方法について説明する。ここで、第１の従来例と、本実施形態とを比較するために、第１の従来例で用いた表記法と同一の表記法を用いることにする。

【００１５】まず、相互情報量を用いたクラスタリングの方法について述べる。ここで、単語数Ｔのテキスト、語数Ｖの語彙、それに語彙の分割関数πとが存在すると仮定し、ここで、語彙の分割関数πは語彙Ｖから語彙の中の単語クラスセットＣへの分割写像（マッピング）を表わす写像関数である。第１の従来例においては、複数の単語からなるテキストデータを生成するバイグラムのクラスモデルの尤度Ｌ（π）は次式によって得られる。

【００１６】

【数１】$L(\pi) = -H + I$

【００１７】ここで、Ｈはモノグラムの単語分布のエントロピーであり、Ｉはテキストデータ内の隣接する２つのクラスC_1，C_2に関する平均的な相互情報量（Average Mutual Information；以下、平均相互情報量とし、ＡＭＩと表記する。）であり、次式で計算することができる。

【００１８】

【数２】

$$I = \sum_{C_1, C_2} Pr(C_1, C_2)\log\{Pr(C_1 \mid C_2)/Pr(C_1)\}$$

$$= \sum_{C_1, C_2} Pr(C_1, C_2)\log\{Pr(C_1, C_2)/(Pr(C_1)\cdot Pr(C_2)\}$$

【００１９】ここで、Ｐｒ（C_1）は第１のクラスC_1の単語の出現確率であり、Ｐｒ（C_2）は第２のクラスC_2の単語の出現確率であり、Ｐｒ（$C_1 \mid C_2$）は、第２のクラスC_2の単語は出現した後に、第１のクラスC_1の単語が出現する条件付き確率であり、Ｐｒ（C_1，C_2）は第１のクラスC_1の単語と第２のクラスC_2の単語が隣接して出現する確率である。従って、上記数２で表されるＡＭＩは、互いに異なる第１のクラスC_1の単語と第２のクラスC_2の単語とが隣接して出現する確率を、上記第１のクラスC_1の単語の出現確率と第２のクラスC_2の単語の出現確率との積で割った相対的な頻度の割合を表わす。

【００２０】エントロピーＨは写像関数πに依存しない値であることから、ＡＭＩを最大にする写像関数は同時にテキストの尤度Ｌ（π）も最大にする。従って、ＡＭＩを単語のクラス構成における目的関数として使用することができる。

【００２１】第１の従来例の相互情報量を用いたクラスタリング方法では、下側層から上側層へのボトムアップのマージ手順を用いている。初期の段階では、各単語をそれぞれ１つのクラスに割り当てる。次いで、すべてのクラスのペア（対）の中で最小のＡＭＩの減少量を与える２つのクラスのペアを探索し、その２つのクラスのペアをマージし、マージ後のクラス数が予め決められた数ｃになるまで上記マージの処理を繰り返す。この第１の従来例の基本的な方法において、例えばコンピュータによって実行される演算時間のコンプレキシティー（又は演算時間のコスト）は、当該処理を以下に示すように直接的に実行したとき、V^5（語彙の語数Ｖの５乗）に比例するオーダーであり、これをＯ（V^5）と表記する。ここで、演算時間のコンプレキシティーは、演算時間がどれぐらいかかるかを示す指標である。

【００２２】＜ステップＡ１＞マージ処理の回数は合計で（Ｖ－ｃ）回であり、このときの演算時間のコンプレキシティーは語彙の語数Ｖに比例するオーダーＯ（Ｖ）である。

＜ステップＡ２＞ｎ回のマージ処理の後には、（Ｖ－ｎ）個のクラスが残り、次のマージ処理の段階では、組み合わせ数$_{V-n}C_2$（すなわち、Ｖ－ｎ個のクラスから２つのクラスをとるときの組み合わせ数）個のマージ処理のテスト又はトライアル（trial）（以下、トライアルといい、ここで、複数回のマージ処理を実行するが、実際にマージして単語分類結果に反映させるのは、このうちの１つであるので、本実施形態ではこのように呼ぶ。）を実行して探索する必要がある。そのうちの１つのみが後のマージ処理で有効化される。従って、このときの演算時間のコンプレキシティーは語彙の語数の２乗V^2に比例するオーダーＯ（V^2）である。

＜ステップＡ３＞第ｎ段階での１つのマージ処理のトライアルには、上記数２を用いてＡＭＩを演算するための（Ｖ－ｎ）^{2}個の項又はクラスに関する加算演算を含む。従って、このときの演算時間のコンプレキシティーは語彙の語数の２乗V^2に比例するオーダーＯ（V^2）である。

【００２３】従って、例えばコンピュータによって実行される全体の演算時間のコンプレキシティーは、語数の５乗V^5に比例するオーダーＯ（V^5）となる。しかしながら、後述するように、冗長的な計算を除くことによって、演算時間のコンプレキシティーを、語彙の語数の３乗V^3に比例するオーダーＯ（V^3）に減らすことも可能である。つまり、次のような本発明に係る方法によれば、上記ステップＡ３の部分を一定時間で実行することができる。

【００２４】＜ステップＢ１＞上記数２は、前の段階のマージ処理で値が変更されたクラスのみについて計算する。従って、演算時間のコンプレキシティーは、第１の従来例におけるオーダーＯ（V^2）から、オーダーＯ（Ｖ）となる。

＜ステップＢ２＞前の段階のマージ処理におけるすべてのトライアルの結果を格納する。従って、演算時間のコンプレキシティーは、第１の従来例におけるオーダーＯ（Ｖ）から、語彙の語数Ｖに依存しない一定のオーダーＯ（１）となる。

【００２５】例えば、Ｖ個のクラス数の語彙から始めて既に（Ｖ－ｋ）回のマージ処理を実行して、ｋ個のクラスC_k（１），C_k（２），…，C_k（ｋ）が残っていると仮定する。この段階でのＡＭＩＩｋは次式で計算される。

【００２６】

【数３】

$$I_k = \sum_{l,m} q_k(l, m)$$

【数４】$q_k(l, m) = p_k(l, m) \log [p_k(l, m) / \{pl_k(l) \, pr_k(m)\}]$

【００２７】ここで、$p_k(l, m)$は、クラスC_k（ｌ）における単語の次に、クラスC_k（ｍ）における単語が続く確率であり、次式のように表される。なお、本明細書及び図面において、表示を明確にするために、ｌ（小文字のエル）としてｌをも用い、ｌ＝ｌとする。

【００２８】

【数５】

$p_k(l, m) = Pr(C_k(l), C_k(m))$

【数６】

$$pl_k(l) = \sum_m p_k(l, m)$$

$$pr_k(m) = \sum_l p_k(l, m)$$

【００２９】上記数３においては、q_kは（ｋ×ｋ）クラスのバイグラム平面テーブルの全体にわたって加算され、ここで、（ｌ，ｍ）のセルはq_k（ｌ，ｍ）で表わす。いま、クラスC_k（ｉ）とクラスC_k（ｊ）とのマージ処理のトライアルを探索したとき、当該マージ処理によるＡＭＩの減少量を、$L_k(i, j) \equiv I_k - I_k(i, j)$とし、ここで、$I_k$（ｉ，ｊ）は当該のマージ処理後のＡＭＩである。

【００３０】図３は、本発明に係る単語分類処理における加算領域及び加減算処理を示すクラスバイグラム平面テーブルの図である。ここで、図３及び、以下に示す図４と図５は、２つのクラスのバイグラムの平面を示す。図３に示すように、上記数３における加算領域Ｐ０は、図３の部分領域Ｐ１、Ｐ２及びＰ３の和から部分領域Ｐ４を減じた部分として表わすことができる。この４つの部分Ｐ１，Ｐ２，Ｐ３，Ｐ４のうち、部分領域Ｐ１の加算値はC_k（ｉ）とC_k（ｊ）とのマージ処理によって変化することはない。従って、ＡＭＩの減少量L_k（ｉ，ｊ）を算出する場合、加算領域Ｐ０を、２次元の領域（すなわち、正方形の領域）から１次元の領域（すなわち、複数のライン又は線）に減らすことが可能である。よって、上記ステップＡ３における演算時間のコンプレキシティーは、オーダーＯ（V^2）からオーダーＯ（Ｖ）に減少させることができる。クラスC_k（ｉ）とクラスC_k（ｊ）とのマージ処理によって生成されるクラスを表わす表記法として、C_k（ｉ＋ｊ）を使用すると、ＡＭＩの減少量L_k（ｉ，ｊ）は次式によって与えられる。

【００３１】

【数７】

$$L_k(i, j) = s_k(i) + s_k(j) - q_k(i, j) - q_k(j, i) - \left(\sum_{l \neq i, j} q_k(l, i+j) + \sum_{m \neq i, j} q_k(i+j, m) + q_k(i+j, i+j) \right)$$

ここで、

【数８】

$$s_k(i) = \sum_l q_k(l, i) + \sum_m q_k(i, m) - q_k(i, i)$$

【００３２】すべてのクラスのペアのＡＭＩの減少量L_kを算出したら、当該ＡＭＩの減少量L_kが最小となるようなペア、例えばクラスC_k（ｉ）とクラスC_k（ｊ）（但しｉ＜ｊ）とを選択し、次いで、これらのクラスのペアをマージさせたときの新しいクラスの名前をC_{k-1}（ｉ）と命名し、さらに、（ｋ－１）個のクラスの新たなセットによる次のマージ処理を続けて実行する。クラスC_k（ｉ）とクラスC_k（ｊ）を除くすべてのクラスについてマージ処理後に同じ方法で索引番号（インデックス）を付与する。すなわち、クラスC_k（ｍ）をクラスC_{k-1}（ｍ）とし、ただし、ｍ≠ｉ，ｊである。ここで、ｊ≠ｋであればクラスC_k（ｋ）をクラスC_{k-1}（ｊ）とし、ｊ＝ｋであればマージ処理後にC_k（ｋ）を削除する。

【００３３】前の段階のマージ処理によるすべてのＡＭＩの減少量L_kを記憶装置に格納することによって、別の最適化処理を実行することができる。ここで、クラスのペア（C_k（ｉ），C_k（ｊ））がマージ処理の対象として選択され、すなわち、すべてのペア（ｌ，ｍ）に対して、$L_k(i, j) \leqq L_k(l, m)$であると仮定する。次のマージ処理の段階では、すべての（ｌ，ｍ）のペアに対して、ＡＭＩの減少量$L_{k-1}^{(i,j)}(l, m)$を計算する必要がある。ここで、上付き文字（ｉ，ｊ）は、クラスのペア（C_k（ｉ），C_k（ｊ））が前のマージ処理の段階でマージされたことを意味している。ここ

で、ＡＭＩの減少量$L_{k-1}^{(i,j)}$（ｌ，ｍ）とL_k（ｌ，ｍ）の違いに注意する必要がある。すなわち、$L_{k-1}^{(i,j)}$（ｌ，ｍ）はクラスｉとクラスｊとのマージ処理の後にクラスｌとクラスｍとをマージしたことによるＡＭＩの減少量であり、L_k（ｌ，ｍ）はクラスｉとクラスｊとのマージ処理なしにクラスｌとクラスｍとをマージしたことによるＡＭＩ減少量である。従って、ＡＭＩの減少量$L_{k-1}^{(i,j)}$（ｌ，ｍ）と、ＡＭＩの減少量L_k（ｌ，ｍ）との差分は、クラスのペア（C_k（ｉ），C_k（ｊ））のマージ処理によって影響を受ける項又はクラスのみから発生する。

【００３４】上記の処理を、図４を参照して説明すると、ＡＭＩの減少量$L_{k-1}^{(i,j)}$（ｌ，ｍ）と、ＡＭＩの減少量L_k（ｌ，ｍ）に対するクラスバイグラム平面テーブルの加算領域は図４の（ｂ）及び（ａ）のようになる。領域｛（ｘ，ｙ）｜ｘ≠ｉ，ｊ，ｌ，ｍ、及びｙ≠ｉ，ｊ，ｌ，ｍ｝の加算値は、クラスｉとクラスｊとのマージ処理によって、あるいは、クラスｌとクラスｍとのマージ処理によって変化することはないため、それらの領域については図示していない。さらに、詳細後述するように、｛$L_{k-1}^{(i,j)}$（ｌ，ｍ）－L_k（ｌ，ｍ）｝を計算するときに、図４の図中のほとんどの領域は互いに相殺されて数カ所のポイントの領域のみが残る。こうして、上記ステップＡ３における演算時間のコンプレキシティーを定数にまで減少することができる。

【００３５】

【数９】$L_k(l,m)=I_k-I_k(l,m)$

及び

$L_{k-1}^{(i,j)}(l,m)=I_{k-1}^{(i,j)}-I_{k-1}^{(i,j)}(l,m)$，

であるので、

$L_{k-1}^{(i,j)}(l,m)-L_k(l,m)=-(I_{k-1}^{(i,j)}(l,m)-I_k(l,m))+(I_{k-1}^{(i,j)}-I_k)$

【００３６】ＡＭＩの減少量$I_{k-1}^{(i,j)}$（ｌ，ｍ）と、ＡＭＩの減少量I_kの加算領域のうちの幾つかの部分は、ＡＭＩの減少量$I_{k-1}^{(i,j)}$の一部、あるいはＡＭＩの減少量I_k（ｌ，ｍ）の一部とともに相殺される。ここで、$Ih_{k-1}^{(i,j)}$（ｌ，ｍ）、Ih_k（ｌ，ｍ）、$Ih_{k-1}^{(i,j)}$、Ih_kはそれぞれ、相殺可能な共通のクラスをすべて相殺した後のＡＭＩの減少量$I_{k-1}^{(i,j)}$（ｌ，ｍ）、I_k（ｌ，ｍ）、$I_{k-1}^{(i,j)}$、I_kであることを表わす。このとき、次のような関係式が与えられる。

【００３７】

【数１０】$L_{k-1}^{(i,j)}(l,m)-L_k(l,m)=-(Ih_{k-1}^{(i,j)}(l,m)-Ih_k(l,m))+(Ih_{k-1}^{(i,j)}-Ih_k)$

ここで、

【数１１】$Ih_{k-1}^{(i,j)}(l,m)=q_{k-1}(l+m,i)+q_{k-1}(i,l+m)$

【数１２】$Ih_k(l,m)=q_k(l+m,i)+q_k(i,l+m)+q_k(l+m,j)+q_k(j,l+m)$

【数１３】$Ih_{k-1}^{(i,j)}=q_{k-1}(i,l)+q_{k-1}(i,m)+q_{k-1}(l,i)+q_{k-1}(m,i)$

【数１４】$Ihk=q_k(i,l)+q_k(i,m)+q_k(j,l)+q_k(j,m)+q_k(l,i)+q_k(l,j)+q_k(m,i)+q_k(m,j)$

【００３８】上記数１０における Ｉｈの加算領域を図５に示す。第１の従来例においては、上記数１０の右辺第２項を無視して第１項のみを使用して、ＡＭＩの減少量$L_{k-1}^{(i,j)}$（ｌ，ｍ）－L_k（ｌ，ｍ）を計算しているようである。なお、上記数１０の第１項に対応する従来技術文献における方程式（１７）の第３式において、符号は正負逆である。しかしながら、上記数１０の第２項は、その第１項と同じ重み係数を有するので、本発明者は、本発明の当該モデルを完全なものとするために、上記数１０を用いる。

【００３９】演算時間のコンプレキシティーのオーダーＯ（V^3）を有する方法を使用する場合でも、語彙数が10^4又はそれ以上のオーダーのように大きいときには、実際に計算することができない。何れにしても、上記ステップＡ１において、オーダーＯ（Ｖ）の演算時間が必要であるため、修正できるのは上記ステップＡ２しかないと考えられる。上記ステップＡ２においては、可能なクラスペアのマージのすべてについて検討することもできるが、実際には探索するクラスペアの範囲を限定することは可能である。このことに関しては、第１の従来例においては、以下のような方法を提案しており、本発明の方法もこれを採用している。まず、互いに重複しない単語数Ｖを含むテキストデータ内の単語に基づいて、Ｖ個の単体のクラスを作り、これを頻度の高い順に配列して、「マージ領域」（本実施形態では、ウィンドウという。従って、本明細書においては、マージ領域とウィンドウとは同義語である。）を、クラス順位の最初のｃ＋１個のクラスの単語とする。従って、まずは、（ｃ＋１）個の頻度の高い単語がマージ領域となる。次いで下記の処理を行う。

【００４０】＜ステップＤ１＞マージ領域内のすべてのペアの中でも、ＡＭＩの減少量を最小にするようなクラスのペアをマージする。

＜ステップＤ２＞（ｃ＋２）番目の位置にあるクラスをマージ領域又はウィンドウの中に挿入し、（ｃ＋２）番目の位置のクラスよりも後ろの各クラスをその左側方向に１つだけ移動させる。

＜ステップＤ３＞残りのクラスが所定のｃ個になるまで上記ステップＤ１とＤ２の処理を繰り返す。

【００４１】当該第１の従来例の処理のアルゴリズムにおいては、上記ステップＡ２の演算時間のコンプレキシティーは、最終クラス数ｃの２乗であるc^2に比例するオーダーＯ（c^2）となり、全体の演算時間は、c^2Vに比例するオーダー（c^2V）に減少する。

【００４２】次いで、単語のクラスタリング構造を得るための方法について述べる。単語のクラスタリング構造を表わすツリーでの表現を得るための最も簡単な方法は、マージ処理における副産物としてデンドログラム（ｄｅｎｄｒｏｇｒａｍ；ツリーの系統図又はツリー図ともいう。）系統樹を構築すること、即ち具体的には、マージの順序の記録（又は履歴）を取ってその記録に基づいて二分木を作ることである。図６に、５単語から成る語彙を使用した簡単な例を示す。図６におけるマージ履歴（又はマージのヒストリともいう。）は、次の表１に示す通りである。なお、表１の第１行目は、「クラスＡとクラスＢとをマージして、マージ後の新しいクラスをＡと名づけた。」ということを意味する。

【００４３】

【表１】

マージ履歴

Ｍｅｒｇｅ（Ａ，Ｂ→Ａ）
Ｍｅｒｇｅ（Ｃ，Ｄ→Ｃ）
Ｍｅｒｇｅ（Ｃ，Ｅ→Ｃ）
Ｍｅｒｇｅ（Ａ，Ｃ→Ａ）

【００４４】しかしながら、この方法を、上記第１の従来例の方法のＯ（C^2V）アルゴリズムに直接的に適用した場合、各クラスのバランスは極端に悪くなり、図７に示されているようなほぼ左側方向の分岐のみのツリー構造となる。この理由は、ＡＭＩ量に関して言えば、マージ領域にある複数のクラスをある一定の大きさを有するように成長させた後に、比較的大きなサイズを有するより高い頻度を有するクラスをマージするよりは、より低い頻度を有する単集合のクラスをマージした方が、大幅にコストが安くなるからである。

【００４５】本発明者が採用した本発明に係る新しい方法は以下の通りである。

＜ステップＥ１＞ＭＩクラスタリング：マージ領域の制約条件を有する相互情報クラスタリングアルゴリズムを使用してｃ個のクラスを作成する。当該ｃ個のクラスは、図１９に示すように、最後のツリー図であるデンドログラムの中間層１００を構成する。

＜ステップＥ２＞外部クラスタリング：テキストデータ中のすべての単語をクラス・トークン（ｃｌａｓｓ　ｔｏｋｅｎ）と置換し（実際には、テキストデータ中の全体の文章の代わりにバイグラム・テーブルについてのみ処理を行う。）、マージ領域の制約条件なしにすべてのクラスがマージ処理によって単一のクラスになるまで二分木の形式でマージ処理（バイナリーマージ処理）を実行する。当該処理によって、デンドログラムD_{root}を作成する。このデンドログラムD_{root}は、例えば図１９に示すように、最終のツリー構造の上側層１０１を構成する。

＜ステップＥ３＞内部クラスタリング：｛C^1，C^2，…，C^i，…，C^c｝を上記ステップＥ１で得られた中間層１００のクラスの集合（クラスセット）とする。そして、それぞれのｉ（$1 \leq i \leq c$；ｉは自然数である。）について以下の処理を行う。

【００４６】＜ステップＥ３－１＞クラスC^iのものを除いて、テキスト中のすべての単語をそのクラス・トークンと置き換える。新しい語彙$V' = V_1 \cup V_2$を決定する。ここで、V_1＝｛C^iにおけるすべての単語｝、V_2＝｛C^1，C^2，…，C^{i-1}，C^{i+1}，C^c｝であり、C^jはｊ番目のクラスのトークンである（$1 \leq j \leq c$）。語彙Ｖ’の各要素を各々のクラスに割り当て、語彙V_1の要素のみを含むクラスに限ってマージ処理が可能となるという制約条件付きマージ処理によって二分木の形式でマージ処理を実行する。当該処理は、最初の$|V_1|$個のクラスにおける語彙V_1の要素（すなわち、単語）を含む語彙Ｖ’の要素を頻度の順序で順序づけし、次いで、最初に幅$|V_1|$を有しかつ各マージ処理によって１つずつ減少する幅を有するマージ領域内でマージ処理を実行することによって実行することができる。ここで、$|V_1|$は語彙V_1の単語の個数である。

＜ステップＥ３－２＞語彙V_1におけるすべての要素が単一のクラスに入るまでマージ処理を繰り返す。各クラス毎に、マージ処理によって、図１９に示すように、下側層１０２のデンドログラムD_{sub}を作成する。このデンドログラムは、葉のノード（ｌｅａｆ　ｎｏｄｅ）がクラス内の各単語を表している各クラスのサブツリーを構成する。

【００４７】＜ステップＥ４＞上側層１０１のデンドログラムD_{root}の各葉のノードを、対応する下側層１０２のデンドログラムD_{sub}と置き換えすることによってデンドログラムを合成し、これによって、全体のデンドログラムを得ることができる。

【００４８】本発明に係る単語クラスタリングの方法は、意味又は統語的特徴が似通った単語が近接した位置に配置された点で、バランスが取れた二分木の形式を有するツリー構造を生成することができる。図８は、本発明の方法を用いて、ウォール・ストリート・ジャーナル（以下、ＷＳＪという。）のコーパスの中で最も使用頻度の高い上位７０，０００語に関して組み立てた５００クラスの内の１クラスに対する下側層のデンドログラムD_{sub}の一例を示したものである。最後に、根のノード（ルートノード（ｒｏｏｔ　ｎｏｄｅ））から葉のノード（リーフノード（ｌｅａｆ　ｎｏｄｅ）に至るパスの追跡し、左側方向の分岐又は右側方向の分岐をそれぞれ表わす０又は１の１ビットを各分岐に割り当てることによって、語彙の中の各単語に対して、ビットストリング（単語ビット）を割り当てることができる。

【００４９】図１０は、図１の単語分類処理部２０の構成を示すブロック図である。図１０を参照して、単語分

類処理部２０の構成及び動作について説明する。図１０において、単語分類処理部２０は、ＣＰＵ５０を備えたコントローラであって、ＣＰＵ５０と、ＣＰＵ５０によって実行される単語分類処理のプログラム及び当該プログラムを実行するために必要なデータを格納するためのＲＯＭ５１と、上記単語分類処理を実行するときに必要なワークエリアであるワークＲＡＭ５２と、上記単語分類処理を実行するときに必要な複数のメモリエリアを有するＲＡＭ５３と、２つのメモリインターフェース５４，５５とを備え、これらの各回路５０乃至５５はバス５６を介して互いに接続される。ここで、メモリインターフェース５４は、テキストデータメモリ１０とＣＰＵ５０との間に設けられ、テキストデータメモリ１０とＣＰＵ５０との間の信号変換などのインターフェース処理を実行するためのインターフェース回路である一方、メモリインターフェース５５は、単語辞書メモリ１１とＣＰＵ５０との間に設けられ、単語辞書メモリ１１とＣＰＵ５０との間の信号変換などのインターフェース処理を実行するためのインターフェース回路である。ＲＡＭ５３は、次のように区分された複数のメモリ部を備える。

【００５０】（ａ）初期化クラス単語メモリ６１：後述する初期化処理によって得られたｖ個の単語及びそのクラスを格納する；

（ｂ）ＡＭＩメモリ６２：後述する中間層クラスタリング処理、上側層クラスタリング処理及び下側層クラスタリング処理において１つのウィンドウ内のクラスの単語の中ですべての組わせの仮ペアを作り、各仮ペアをマージしたときの平均相互情報量を数２を用いて計算した結果を格納する；

（ｃ）中間層メモリ６３：後述する中間層クラスタリング処理によって得られたｃ個の中間層のクラスの単語を格納する；

（ｄ）上側層ヒストリメモリ６４：後述する上側層クラスタリング処理における各マージ処理の履歴（又はヒストリ）を格納する；

（ｅ）上側層ツリーメモリ６５：上記上側層クラスタリング処理によって得られたツリー図であるデンドログラムD_{root}を格納する；

（ｆ）下側層ヒストリメモリ６６：上記下側層クラスタリング処理によって得られた、中間層１００の各クラスに対して１つのツリー図である複数ｃ個のデンドログラムD_{sub}を格納する；

（ｇ）下側層ツリーメモリ６７：上記下側層クラスタリング処理によって得られたツリー図であるデンドログラムD_{sub}を格納する；

（ｈ）ツリーメモリ６７：上側層１０１の１つのデンドログラムと下側層１０２の複数ｃ個のデンドログラムとを、中間層１００を介して連結することにより得られた全体のツリー図であるデンドログラムを格納する。

【００５１】図１１は、図１の単語分類処理部２０によって実行されるメインルーチンの単語分類処理を示すフローチャートである。図１１に示すように、まず、ステップＳ１においてテキストデータに基づいて出現頻度の高い単語から順に並べる処理を実行する初期化処理を実行し、次いで、ステップＳ２において中間層１００のクラスの単語を求める中間層クラスタリング処理を実行し、さらに、ステップＳ３において上側層１０１のツリー図を求める上側層クラスタリング処理を実行し、そして、ステップＳ４において下側層１０２のツリー図を求める下側層クラスタリング処理を実行し、最後に、ステップＳ５において上側層１０１の１つのツリー図と下側層１０２の複数ｃ個のツリー図とを、中間層１００を介して連結することにより得られた全体のツリー図であるデンドログラムを求めて、その結果を単語辞書として単語辞書メモリ１１に格納するデータ出力処理を実行する。これによって、単語分類処理が終了する。なお、これらのツリー図においては、各単語がそれぞれ１つのクラスに分類されかつクラス間の連結関係が示される。

【００５２】図１２は、図１１のサブルーチンの初期化処理（Ｓ１）を示すフローチャートである。図１２に示すように、ステップＳ１１において、テキストデータメモリ１０内のテキストデータに基づいて、単語の重複を省いたすべての複数ｖ個の単語の出現頻度を調べて、出現頻度の高い単語から順に並べて、これを複数ｖ個のクラスに割り当てて、複数ｖ個のクラスの単語を初期化クラス単語メモリ６１に記憶して、元のメインルーチンに戻る。ここで、ｖは２以上の自然数である。

【００５３】図１３は、図１１のサブルーチンの中間層クラスタリング処理（Ｓ２）を示すフローチャートである。図１３に示すように、まず、ステップＳ２１において、初期化クラス単語メモリ６１から複数ｖ個のクラスの単語を読み出した後、複数ｖ個のクラスの単語のうちの出現頻度の高いクラスの単語からｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ（又はマージ領域）内のクラスの単語として、図１７に示すように、ワークＲＡＭ５２に記憶する。ここで、$1<c+1<v$である。次いで、ステップＳ２２において、ワークＲＡＭ５２に記憶された１つのウィンドウ内のクラスの単語の中で、すべての２個ずつの組み合わせの仮ペアを作り、各仮ペアをそれぞれマージしたときの平均相互情報量を数２を用いて計算して、各仮ペアとそれに対応する計算された平均相互情報量とを次の表２の形式でＡＭＩメモリ６２に記憶する。

【００５４】

【表２】

仮ペア	平均相互情報量
(C^1, C^2)	０．８６７６７８
(C^2, C^3)	０．２３４６８９

(C^3, C^4)	０．１２５６８６
.....	.....
(C^c, C^{c+1})	０．６７５６４２

【００５５】次いで、ステップＳ２３において、図１８に示すように、ＡＭＩメモリ６２に記憶された各仮ペアの平均相互情報量のうち、最大となる仮ペアを見つけて当該仮ペアをマージすることにより、１つのクラスが減少し、マージ後の１つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークＲＡＭ５２に記憶する。そして、ステップＳ２４において、ウィンドウ外のクラスはなくなりかつウィンドウ内のクラスの数はｃ個となったか否かが判断され、その判断がＮＯであるとき、ステップＳ２５において、図１８に示すように、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語をウィンドウ内に挿入し、挿入後の１つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークＲＡＭ５２に記憶した後、ステップＳ２２に戻って、ステップＳ２２以降の処理を繰り返す。

【００５６】一方、ステップＳ２４においてＹＥＳであるときは、ステップＳ２６において、ワークＲＡＭ５２に記憶された、ウィンドウ内のｃ個のクラス及びそれに属する単語を中間層１００として中間層メモリ６３に記憶する。これによって、中間層クラスタリング処理が終了し、メインルーチンに戻る。

【００５７】図１４は、図１１のサブルーチンの上側層クラスタリング処理（Ｓ３）を示すフローチャートであり、図１９に示すように、中間層１００から矢印２０１の方向でツリー図を求める処理である。図１４に示すように、まず、ステップＳ３１において、中間層メモリ６３内のｃ個のクラスの単語を読み出した後、当該ｃ個のクラスの単語を１つのウィンドウ内のクラス単語として、ワークＲＡＭ５２に記憶する。次いで、ステップＳ３２において、ステップＳ２２と同様に、ワークＲＡＭ５２に記憶された１つのウィンドウ内のクラスの単語の中で、すべての２個ずつの組み合わせの仮ペアを作り、各仮ペアをそれぞれマージしたときの平均相互情報量を数２を用いて計算して、各仮ペアとそれに対応する計算された平均相互情報量とを前述の表２の形式でＡＭＩメモリ６２に記憶する。

【００５８】次いで、ステップＳ３３において、ステップＳ２３と同様に、ＡＭＩメモリ６２に記憶された各仮ペアの平均相互情報量のうち、最大となる仮ペアを見つけて当該仮ペアをマージすることにより、１つのクラスが減少し、マージ後の１つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークＲＡＭ５２に記憶する。また、例えば表１の形式を有し、どのクラスとどのクラスとがマージされて新しく名づけられたクラスとなったかを表わす当該マージ処理の履歴を上側層ヒストリメモリ６４に記憶する。そして、ステップＳ３４において、ウィンドウ内のクラスの数はｃ個となったか否かが判断され、その判断がＮＯであるとき、ステップＳ３２に戻って、ステップＳ３２以降の処理を繰り返す。

【００５９】一方、ステップＳ３４においてＹＥＳであるときは、ステップＳ３５において、上側層ヒストリメモリ６４内の上側層の履歴又はヒストリに基づいて、例えば図６に示すように、上側層のツリー図又はデンドログラムD_{root}を作成して上側層ツリーメモリ６５に記憶する。これによって、上側層クラスタリング処理が終了し、メインルーチンに戻る。

【００６０】図１５は、図１１のサブルーチンの下側層クラスタリング処理（Ｓ４）を示すフローチャートであり、図１５に示すように、下側層１０２の底辺に位置する単語から、中間層１００の各クラスC^i毎に、矢印２０２の方向でツリー図を求める処理である。ある。図１５に示すように、まず、ステップＳ４１において、中間層メモリ６３内のｃ個のクラスの単語を読み出した後、当該ｃ個のクラスから１つのクラスを選択する。そして、ステップＳ４２において、選択されたクラス内のv_i個の単語を１つのウィンドウ内のクラス単語として、ワークＲＡＭ５２に記憶する。次いで、ステップＳ４３において、ステップＳ２２及びＳ３２と同様に、ワークＲＡＭ５２に記憶された１つのウィンドウ内のクラスの単語の中で、すべての２個ずつの組み合わせの仮ペアを作り、各仮ペアをそれぞれマージしたときの平均相互情報量を数２を用いて計算して、各仮ペアとそれに対応する計算された平均相互情報量とを前述の表２の形式でＡＭＩメモリ６２に記憶する。

【００６１】次いで、ステップＳ４４において、ステップＳ２３及びＳ３３と同様に、ＡＭＩメモリ６２に記憶された各仮ペアの平均相互情報量のうち、最大となる仮ペアを見つけて当該仮ペアをマージすることにより、１つのクラスが減少し、マージ後の１つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークＲＡＭ５２に記憶する。また、例えば表１の形式を有し、どのクラスとどのクラスとがマージされて新しく名づけられたクラスとなったかを表わす当該マージ処理の履歴を下側層ヒストリメモリ６６に記憶する。そして、ステップＳ４５において、ウィンドウ内のクラスの数はｃ個となったか否かが判断され、その判断がＮＯであるとき、ステップＳ４３に戻って、ステップＳ４３以降の処理を繰り返す。ここで、ステップＳ４３及びＳ４４の処理は、中間層１００の各クラス毎に実行される。

【００６２】一方、ステップＳ４５においてＹＥＳであるときは、ステップＳ４６においてすべての中間層１００のクラスについて処理したか否かが判断され、当該判断がＮＯであるとき、未処理のクラスが残っているので、ステップＳ４７において残っている中間層１００の

別の未処理のクラスを選択した後、ステップＳ４２に進む。一方、ステップＳ４６においてＹＥＳであるときは、ステップＳ４８において、下側層ヒストリメモリ６６内の下側層の履歴又はヒストリに基づいて、例えば図６に示すように、下側層のツリー図又はデンドログラムD_{sub}を作成して下側層ツリーメモリ６７に記憶する。これによって、下側層クラスタリング処理が終了し、メインルーチンに戻る。

【００６３】図１６は、図１１のサブルーチンのデータ出力処理（Ｓ５）を示すフローチャートである。図１６に示すように、まず、ステップＳ５１において、図１９に示すように、上側層ツリーメモリ６５内の上側層のツリー図と、下側層ツリーメモリ６７内の下側層のツリー図とに基づいて、これら２つのツリー図を中間層１００の各クラスC^iを介して連結し、すなわち、上側層ツリーメモリ６５内の上側層のツリー図を中間層１００の各クラスC^iに連結する一方、下側層ツリーメモリ６７内の下側層ツリー図をその頂点にあるクラスを中間層１００の各クラスC^iに連結する。これによって、当該テキストデータに基づく全体のツリー図を作成して、ツリー図の情報をツリーメモリ６８に記憶する。当該ツリーメモリ６８には、図６及び図８に示すように、各クラスの単語間の連結関係が単語辞書として記憶される。そして、ステップＳ５２において、ツリーメモリ６８内のツリー図の情報を単語分類結果（又は単語クラスタリング結果）として単語辞書メモリ１１に出力して記憶する。

【００６４】＜第１の実施形態＞図１は、本発明に係る第１の実施形態である音声認識装置のブロック図である。図１において、テキストデータメモリ１０内に格納された、例えば英語又は日本語の複数の単語を含むテキストデータは、単語分類処理部２０によって上述の単語分類処理が実行されることにより、複数のクラスに分類されかつクラスの連結関係が記述された単語辞書として、単語辞書メモリ１１内に格納される。

【００６５】一方、マイクロホン１に入力された複数の単語からなる発声音声は、マイクロホン１によって音声信号に変換された後、Ａ／Ｄ変換器２によってディジタル音声信号にＡ／Ｄ変換される。ディジタル音声信号は特徴抽出部３に入力され、特徴抽出部３は、入力されたディジタル音声信号に対して例えばＬＰＣ分析してケプストラム係数や対数パワーなどの特徴パラメータを抽出して、バッファメモリ４を介して音声認識部５に出力する。音声認識部５は、単語辞書メモリ１１に格納された単語辞書を参照しかつ、例えば音素隠れマルコフモデル（以下、音素HMMという。）である言語モデルメモリ１２に格納された言語モデルを参照して、単語毎に音声認識を実行して、音声認識結果を出力する。なお、ここで、単語辞書メモリ１１内の単語辞書は、例えば、

（ａ）０１００１００１０，ｐｏｓｉｔｉｏｎ；
（ｂ）０１００１００１１，ｌｏｃａｔｉｏｎ；
（ｃ）１１００１０１００，ｆｏｒ；

のように各単語とその単語の属するクラスを表現するビット列などの情報を含む。

【００６６】＜第２の実施形態＞図２は、本発明に係る第２の実施形態である形態素及び構文解析装置のブロック図である。図２において、テキストデータメモリ３１，３２にそれぞれ格納された、複数の単語からなる２つのテキストデータはそれぞれ、単語分類処理部２０によって上述の単語の分類処理が実行されることにより、複数のクラスに分類されかつクラスの連結関係が記述された単語辞書として、それぞれ単語辞書メモリ４１，４２内に格納される。

【００６７】日本語又は英語などの所定の言語の文字列からなり複数の単語からなる自然言語文が形態素解析部２１に入力され、形態素解析部２１は、入力された自然言語文の各単語の出現形に対して、単語辞書メモリ４１に格納された単語辞書を参照して上記自然言語文を複数の単語に分割するとともに、上記各出現形に対して品詞、活用形、標準表現形、及び類語コードなどの情報を付与し、これらの解析結果を構文解析部２２に出力する。次いで、構文解析部２２は、単語辞書メモリ４２に格納された単語辞書を参照して、所定の構文解析を実行して単語列に対して構文木情報を付加して解析結果として出力する。

【００６８】以上説明したように、図１９に示すように、下側層、中間層、及び上側層と階層化して、複数の単語を二分木の形式で複数のクラスに分類したので、単語分類処理によりバランスのとれた階層構造を有する単語分類結果を得ることができる。また、ＡＭＩの計算においては、下側層、中間層、及び上側層ともに、すべてのクラスの単語を対象としてＡＭＩを計算しているので、計算されたＡＭＩは局所的なＡＭＩではなく、全体の単語の情報を含んだグローバルはＡＭＩに基づいて、クラスタリング処理を実行している。従って、全体的に最適化された単語分類結果を得ることができる。これにより、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができる。さらに、上記単語分類部２０により得られた単語辞書に基づいて音声認識することにより、従来例に比較して高い認識率で音声認識することができる。

【００６９】以上の実施形態において、音声認識部５と、単語分類処理部２０と、形態素解析部２１と、構文解析部２２とは例えばディジタル計算機によって構成される。以上の実施形態の単語分類処理は、図１１に示すように、中間層クラスタリング処理、上側層クラスタリング処理、下側層クラスタリング処理の順序で実行しているが、本発明はこれに限らず、中間層クラスタリング処理、下側層クラスタリング処理、上側層クラスタリング処理の順序で実行してもよい。以上の実施形態において、図１１の初期化処理の前に、単語のｎ－グラムを利

用して、同一の単語に隣接する割合の多い単語を同一のクラスに割り当てるという基準で複数の単語を複数のクラスに分類する処理を実行してもよい。

【００７０】

【実施例】

＜実験（シミュレーション）＞６年分のＷＳＪのコーパスの平易なテキストデータを使用し、クラスタと単語ビットを作成した。テキストのサイズは５００万語、１０００万語、２０００万語、及び５０００万語（それぞれ、５ＭＷ、１０ＭＷ、２０ＭＷ、５０ＭＷ；ここで、Ｗはワードである。）である。語彙はコーパス全体で最も頻繁に使用されている上位７万語とした。最終クラス数ｃは５００に設定した。獲得したクラスタと単語ビットを、それぞれ次の２つの尺度ＳＳ１とＳＳ２を使用して評価する。

（ａ）尺度ＳＳ１は、ＷＳＪのコーパス、及び本出願人が所有するコーパスである一般的な英語のツリーバンクに基づいた、クラスを基本としたトライグラムモデルのパープレキシティーを計算するパープレキシティー法である。

（ｂ）尺度ＳＳ２は、本出願人が所有する決定木を用いる部分音声のラベル付け（ｔａｇｇｉｎｇ又はｌａｂｅｌｉｎｇ）のラベル付け装置（ｔａｇｇｅｒ又はｌａｂｅｌｅｒ）における誤り率である。

【００７１】＜パープレキシティ法＞単語をその所属クラスに写像するクラス関数Ｇを使用すると単語トライグラムの確率は、次式のように書き直すことができる。

【００７２】

【数１５】$P(w_i \mid w_{i-2}w_{i-1}) = P_c(G(w_i)) \mid G(w_{i-2})G(w_{i-1}))\, P_m(w_i \mid G(w_i))$

【００７３】ここで、P_cは２次のマルコフ連鎖確率であり、P_mは単語メンバーシップ確率である。Ｐｃ及びＰｍのスムージングは、それぞれカッツ（Ｋａｔｚ）のバックオフ、及びグッドテューリング公式を使用して行う。トレーニング用テキストのサイズは１．９ＭＷで、テストテキストは１５０ＫＷであり、両者ともＷＳＪのコーパスを典拠としている。語彙サイズは７７ＫＷである。図９は、テストテキストのパープレキシティーとクラスタリングのテキストサイズとの関係を示している。クラスタリングのテキストサイズにおけるゼロ点は、単語トライグラムモデルのパープレキシティーを表している。クラスタリングのテキストサイズが増加するに従って、パープレキシティーは単調に減少する。これはクラスタリング処理の改善を示している。５０ＭＷでは、パープレキシティーは単語トライグラムモデルの場合より１８％低くなっている。この結果は、クラス・トライグラムのパープレキシティーが単語トライグラムモデルの場合より僅かに高いとした第１の従来例の結果とは好対照である。

【００７４】＜決定木を用いた音声部分のラベル付け＞本出願人が所有する決定木を用いる部分音声のラベル付け(tagging)のラベル付け装置（tagger）は、スパッター（ＳＰＡＴＴＥＲ、例えば、従来技術文献２「D. Magerman, “Natural Language Parsing as Statistical Recognition”, Doctoral Dissertation, Stanford University, Stanford, California, １９９４年」参照。）をベースとした、本出願人が所有する決定木パーザーの統合モジュールである。上記ラベル付け装置は、ユニバーシティ・オブ・ペンシルバニアのトリーバンクプロジェクトのそれよりも、１桁だけ大きい４４１個の統語的ラベル（syntactic tags）を採用している。学習用テキスト、テスト用テキスト、及び実行用テキストはすべて、単語とラベルとの対のすべてのシーケンスを含む。学習段階では、イベント（event）は、特徴値の集合又は、質問とそれに対する回答との対の集合である。１つの特徴は、処理すべき現在の単語ｗｏｒｄ（０）が現れる文脈における任意の属性であり、これは便宜上、質問の形式で表される。ラベル付けは左から右へと行う。表３は、処理すべき現在の単語“ｌｉｋｅ”を用いたイベントの一例を示している。

【００７５】

【表３】

```
Event－１２８:
{
<word(0), “like”><word(-1), “flies”><word(-2), “time”>
<word(1), “an”><word(2), “arrow”><tag(-1), “Verb-3rd-Sg-type3”
>
<tag(-2), “Noun-Sg-type14”>
 ..........                                  (Basic Questions)
<Inclass?(word(0), Class295), “yes”><WordBits(Word(-1), 29), “1”>
 ..........                                  (WordBits Questions)
<Tag, “Prep-type5”>
}
```

【００７６】このイベントの最後のペアは、回答、即ち　当該現在の単語の正しいラベルを示す特別な項目であ

る。最初の２行は当該現在の単語の回りの単語の識別に関する質問と、先行する単語のためのラベルを表している。これらの質問は、基本質問と呼ばれている。第２の質問形式（単語ビット質問）は、「この現在の単語はクラス２９５にありますか？」或いは「先行する単語の単語ビット中の第２９ビットは何ですか？」と言ったクラスタ及び単語ビットに関するものである。

【００７７】イベントの集合から決定木を作成する。決定木の根のノードは、それぞれ対応する単語に対して正しいラベルを含んでいるすべてのイベントからなるセットを表している。根のノード用のラベルの確率分布は当該集合におけるラベルの相対的な頻度を計算することによって得ることができる。当該セットの中の各イベントにおける特徴値を問い合わせることで、そのセットはＮ個のサブセットに分割することができる（Ｎは特徴に関する可能値である）。次いで、この特徴値を条件として、各サブセットに対するラベルの条件付き確率分布を計算することが可能である。セットの分割によって生じるエントロピーの減少を各特徴毎に計算した後、エントロピーの減少量を最大にする特徴を選択する。この方法を反復し、セットを各サブセットに分割することによって、葉のノードがタグの条件付き確率分布を含むような決定木を構築することができる。次いで、獲得した確率分布を実行用データを使用してスムージングする。スムージング処理の詳細については上記従来技術文献２を参照せよ。テスト段階では、システムはテストテキスト内の各単語に対する条件付き確率分布を調査し、ビームサーチを使用して最も可能性のあるラベル付けシーケンスを選択する。

【００７８】本発明者がラベル付け実験に使用したのは、ＷＳＪのテキスト、及び本出願人が所有するコーパス（以下、ＡＴＲコーパスという。）である。ＷＳＪのテキストは、本出願人の統語ラベルセットを使用して手動でラベル付けをし直した。上記ＡＴＲコーパスは、文語体の米語の包括的な見本であり、その語法のスタイル及び設定は非常に幅広く、多くの異なる領域から作り上げられている。ＡＴＲコーパスはまだ開発過程にあるため、この実験用として手元にあるテキストの大きさは、ラベルセットが大型である割にはかなり小型である。表４は今回の実験に使用したテキストのサイズを示している。

【００７９】

【表４】

テキストサイズ(単語数)	学習用	テスト用	実行用
ＷＳＪのテキスト	７５，１３９	５，８３１	６，５３４
ＡＴＲコーパス	７６，１３２	２３，１６３	６，６８０

【００８０】図２０は、多様なクラスタリングのテキストサイズに対するラベル付けの誤り率を表している。本実験では、２種類の質問形式の中から基本質問及び単語ビット質問を使用している。ラベル付け装置への単語ビット情報の導入の効果を見るため異なる実験を行ったが、その実験では無作為に生成されたビットストリングを各単語に割り当て（特徴的なビットストリングが各単語に割り当てられているが、ラベル付け装置もビットストリングを処理中の各単語の認識番号として使用している。この制御実験においては、ビットストリングの割り当ては無作為に行なわれるが、２つの単語が同じ単語ビットを持つことはない。無作為の単語ビットは、ラベル付け装置に対して単語の認識以外には何のクラス情報も与えない。）、基本質問と単語ビット質問を使用した。結果はクラスタリングのテキストサイズのゼロの値において表されている。ＷＳＪのテキスト及びＡＴＲコーパスの何れも、ラベル付けの誤り率は、５ＭＷのテキストから抽出された単語ビット情報を使用することによって３０％以上低下し、また、クラスタリングのテキストサイズが増加するとさらに誤り率が減少した。５０ＭＷでは、誤り率は４３％も低下した。これもまた、クラスタの品質向上はクラスタリングのテキストサイズの増大によるものであることを示している。全体的にみて、高い誤り率は非常に大きなラベルセットと、小さな学習用セットに起因している。この結果の注目に値する点は、ＡＴＲコーパスのテキストとＷＳＪのテキストは互いに領域が非常に異なったものであるにも関わらず、ＷＳＪのテキストから構成された単語ビットの導入が、ＷＳＪのテキストに対してと同じくらいＡＴＲコーパスのテキストのラベル付けにも効果的であったことである。この点から、獲得した階層的クラスタは領域を越えて移動可能であると考えられる。

【００８１】以上説明したように、本発明者は、複数の単語の階層的クラスタリング分割に関するアルゴリズムを提案し、５ＭＷから５０ＭＷまでの大型テキストデータを使用したクラスタ分割の実験を行った。獲得したクラスタの高品質性は、２種類の評価方法によって確認されている。クラスを基にしたトライグラムモデルのパープレキシティーは、単語をベースとしたトライグラムモデルの場合よりも１８％低くなっている。本出願人が所有する決定木の音声部分のラベル付け装置に単語ビットを導入することにより、ラベル付けの誤りの割合は４３％も減少する。ＷＳＪのテキストから得る階層的クラスタリング分割処理はまた、ＷＳＪのテキストとは全く異

なる範囲にあるＡＴＲコーパスのテキストのラベル付けにも有効であることが判った。

【００８２】

【発明の効果】以上詳述したように本発明に係る請求項１記載の単語分類処理方法によれば、複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当てるステップと、上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１の記憶装置に記憶するステップと、上記第１の記憶装置に記憶された１つのウィンドウ内のクラスの単語に基づいて、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する確率を、上記第１のクラスの単語の出現確率と第２のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数ｃ個のクラスに分類し、分類された複数ｃ個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数ｃ個のクラスとして第２の記憶装置に記憶するステップと、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第３の記憶装置に記憶するステップと、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスの各クラス毎に、上記中間層の複数ｃ個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第４の記憶装置に記憶するステップと、上記第４の記憶装置に記憶された上記ツリー図の下側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結する一方、上記第３の記憶装置に記憶された上記ツリー図の上側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第５の記憶装置に記憶するステップとを備える。従って、下側層、中間層、及び上側層と階層化して、複数の単語を二分木の形式で複数のクラスに分類したので、単語分類処理によりバランスのとれた階層構造を有する単語分類結果を得ることができる。また、平均相互情報量の計算においては、下側層、中間層、及び上側層ともに、すべてのクラスの単語を対象として平均相互情報量を計算しているので、計算された平均相互情報量は局所的な平均相互情報量ではなく、全体の単語の情報を含んだグローバルは平均相互情報量に基づいて、クラスタリング処理を実行している。従って、全体的に最適化された単語分類結果を得ることができる。これにより、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができる。

【００８３】また、請求項２記載の単語分類処理方法は、請求項１記載の単語分類処理方法において、上記分類された複数ｃ個のクラスを上記第２の記憶装置に記憶するステップは、上記第１の記憶装置に記憶された１つのウィンドウよりも外側のクラスが存在し、又は上記１つのウィンドウ内のクラスがｃ個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする。従って、所定の複数ｃ個のクラスを有する中間層を最適化形式で得ることができる。

【００８４】本発明に係る請求項３記載の単語分類処理装置は、複数の単語を含むテキストデータに対して、互いに異なるすべての複数ｖ個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数ｖ個のクラスに割り当てる第１の制御手段と、上記複数ｖ個のクラスの単語のうち出現頻度が高いｖ個未満の（ｃ＋１）個のクラスの単語を１つのウィンドウ内のクラスの単語として第１の記憶装置に記憶する第２の制御手段と、上記第１の記憶装置に記憶された１つのウィンドウ内のクラスの単語に基づいて、互いに異なる第１のクラスの単語と第２のクラスの単語とが隣接して出現する確率を、上記第１のクラスの単語の出現確率と第２のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数ｃ個のクラスに分類し、分類された複数ｃ個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数ｃ個のクラスとして第２の記憶装置に記憶する第３の制御手段と、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第３の記憶装置に記憶する第４の制御手段と、上記第２の記憶装置に記憶された中間層の複数ｃ個のクラスの各クラス毎に、上記中間層の複数ｃ個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で１個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第４の記憶装置に記憶する第５の制御手段と、上記第４の記憶装置に記憶された上記ツリー図の下側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結する一方、上記第３の記憶装置に記憶された上記ツリー図の上側層を、上記第２の記憶装置に記憶された上記中間層の複数ｃ個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として

第５の記憶装置に記憶する第６の制御手段とを備える。従って、下側層、中間層、及び上側層と階層化して、複数の単語を二分木の形式で複数のクラスに分類したので、単語分類処理によりバランスのとれた階層構造を有する単語分類結果を得ることができる。また、平均相互情報量の計算においては、下側層、中間層、及び上側層ともに、すべてのクラスの単語を対象として平均相互情報量を計算しているので、計算された平均相互情報量は局所的な平均相互情報量ではなく、全体の単語の情報を含んだグローバルは平均相互情報量に基づいて、クラスタリング処理を実行している。従って、全体的に最適化された単語分類結果を得ることができる。これにより、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができる。

【００８５】また、請求項４記載の単語分類処理装置は、請求項３記載の単語分類処理装置において、上記第３の制御手段は、上記第１の記憶装置に記憶された１つのウィンドウよりも外側のクラスが存在し、又は上記１つのウィンドウ内のクラスがｃ個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行する。従って、所定の複数ｃ個のクラスを有する中間層を最適化形式で得ることができる。

【００８６】本発明に係る請求項５記載の音声認識装置によれば、入力される発声音声の音声信号に基づいて、請求項３又は４記載の単語分類処理装置によって複数の単語が複数のクラスに分類された単語分類結果を含む単語辞書と、所定の隠れマルコフモデルとを参照して上記発声音声を音声認識する音声認識手段を備える。従って、上記単語分類処理装置により得られた、バランスのとれた階層構造を有しかつ全体的に最適化された単語辞書に基づいて音声認識することにより、従来例に比較して高い認識率で音声認識することができる。

【図面の簡単な説明】

【図１】　本発明に係る第１の実施形態である音声認識装置のブロック図である。

【図２】　本発明に係る第２の実施形態である形態素及び構文解析装置のブロック図である。

【図３】　図１及び図２の単語分類処理部によって実行される単語分類処理における加算領域及び加減算処理を示すクラスバイグラム平面テーブルの図である。

【図４】　（ａ）は上記単語分類処理におけるＡＭＩ減少量Ｌｋ（ｌ，ｍ）に対する加算領域を示すクラスバイグラム平面テーブルの図であり、（ｂ）は上記単語分類処理におけるＡＭＩ減少量$Ｌｋ^{(i,j)}$（ｌ，ｍ）に対する加算領域を示すクラスバイグラム平面テーブルの図である。

【図５】　（ａ）は上記単語分類処理におけるマージ後のＡＭＩ量$Ｉｈ_{k}$を示すクラスバイグラム平面テーブルの図であり、（ｂ）は上記単語分類処理におけるマージ後のＡＭＩ量$Ｉｈ_{k}$（ｌ，ｍ）を示すクラスバイグラム平面テーブルの図であり、（ｃ）は上記単語分類処理におけるマージ後のＡＭＩ量$Ｉｈ_{k-l}^{(i,j)}$を示すクラスバイグラム平面テーブルの図であり、（ｄ）は上記単語分類処理におけるマージ後のＡＭＩ量$Ｉｈ_{k-l}^{(i,j)}$（ｌ，ｍ）を示すクラスバイグラム平面テーブルの図である。

【図６】　上記単語分類処理によって得られるデンドログラム（ツリーの系統図）の一例を示す図である。

【図７】　上記単語分類処理によって得られる左側方向の分岐ツリーの一例を示す図である。

【図８】　上記単語分類処理によって得られる１つのクラスに対するサブツリーの一例を示す図である。

【図９】　本発明の音声認識装置におけるシミュレーション結果である、テキストの大きさに対するパープレキシティーを示すグラフである。

【図１０】　図１の単語分類処理部２０の構成を示すブロック図である。

【図１１】　図１の単語分類処理部２０によって実行されるメインルーチンの単語分類処理を示すフローチャートである。

【図１２】　図１１のサブルーチンの初期化処理（Ｓ１）を示すフローチャートである。

【図１３】　図１１のサブルーチンの中間層クラスタリング処理（Ｓ２）を示すフローチャートである。

【図１４】　図１１のサブルーチンの上側層クラスタリング処理（Ｓ３）を示すフローチャートである。

【図１５】　図１１のサブルーチンの下側層クラスタリング処理（Ｓ４）を示すフローチャートである。

【図１６】　図１１のサブルーチンのデータ出力処理（Ｓ５）を示すフローチャートである。

【図１７】　図１１のサブルーチンの中間層クラスタリング処理（Ｓ２）におけるステップＳ２１の処理を示し、単語クラスの集合を示す図である。

【図１８】　図１１のサブルーチンの中間層クラスタリング処理（Ｓ２）におけるステップＳ２３及びＳ２４の処理を示し、単語クラスの集合を示す図である。

【図１９】　図１１の単語分類処理における処理及びその処理によって得られる階層構造を示す図である。

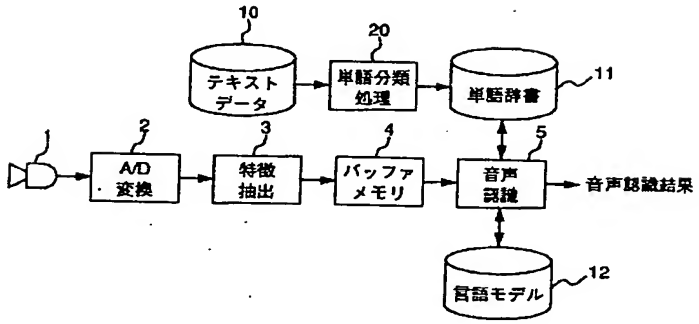
【図２０】　本発明の音声認識装置のシミュレーション結果である、テキストの大きさに対するクラスタリング処理後のラベル付けの誤り率を示すグラフである。

【符号の説明】

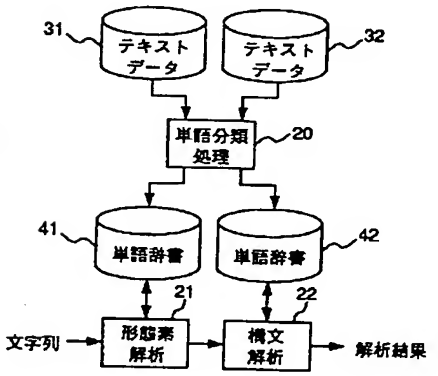
１…マイクロホン、
２…Ａ／Ｄ変換器、
３…特徴抽出部、
４…バッファメモリ、
５…音声認識部、
１０，３１，３２…テキストデータメモリ、

１１，４１，４２…単語辞書メモリ、
１２…言語モデル、
２０…単語分類処理部、
２１…形態素解析部、
２２…構文解析部、
５０…ＣＰＵ、
５１…ＲＯＭ、
５２…ワークＲＡＭ、
５３…ＲＡＭ、
５４，５５…メモリインターフェース、
６１…初期化クラス単語メモリ、
６２…ＡＭＩメモリ、
６３…中間層メモリ、
６４…上側層ヒストリメモリ、
６５…上側層ツリーメモリ、
６６…下側層ヒストリメモリ、
６７…下側層ツリーメモリ、
６８…ツリーメモリ、
１００…中間層、
１０１…上側層、
１０２…下側層、
Ｓ１…初期化処理、
Ｓ２…中間層クラスタリング処理、
Ｓ３…上側層クラスタリング処理、
Ｓ４…下側層クラスタリング処理、
Ｓ５…データ出力処理。

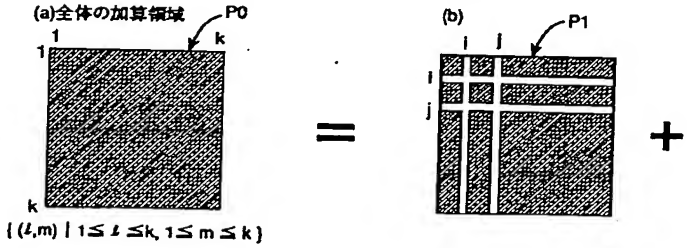
【図１】



【図２】

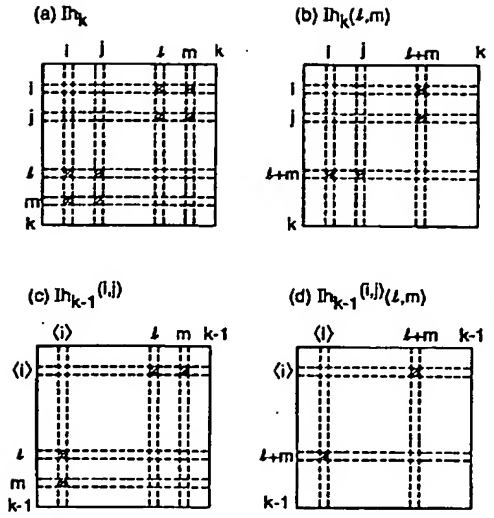


【図３】

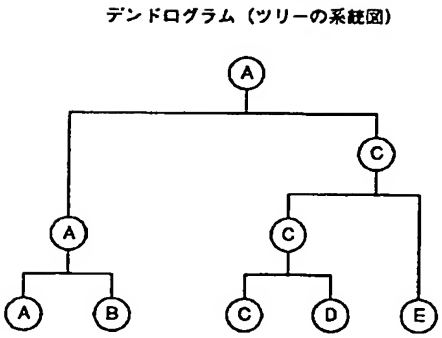


【図４】

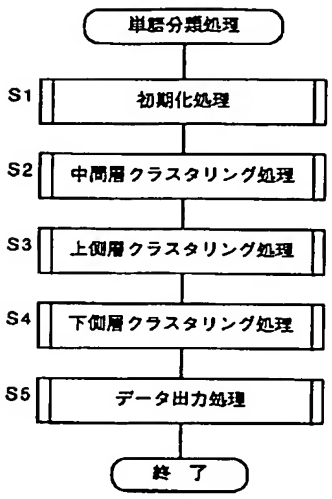
【図５】



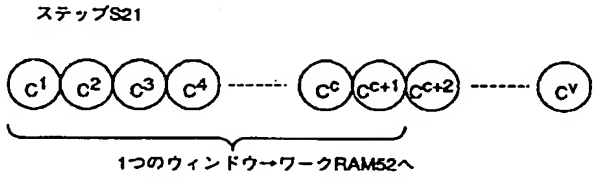
【図６】



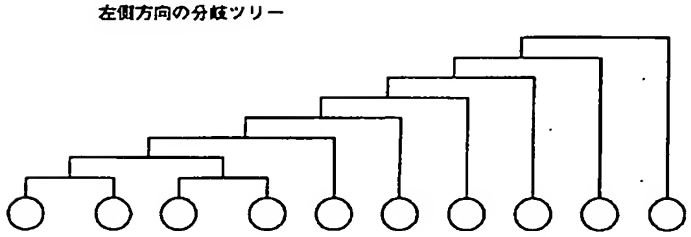
【図１１】



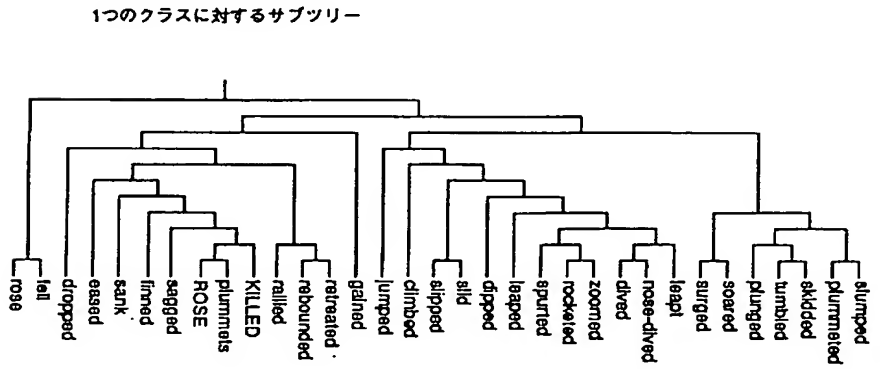
【図１７】



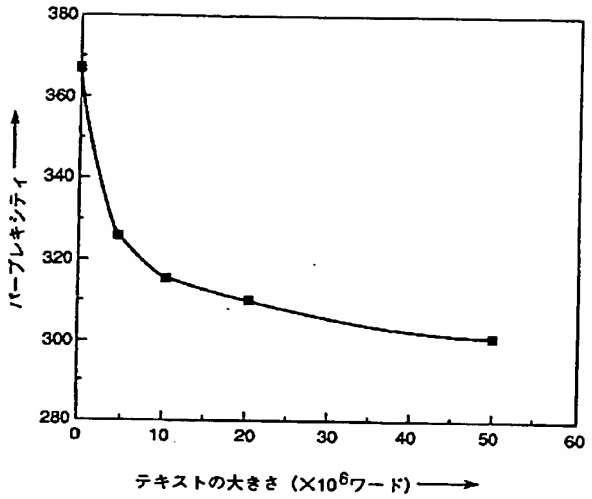
【図７】



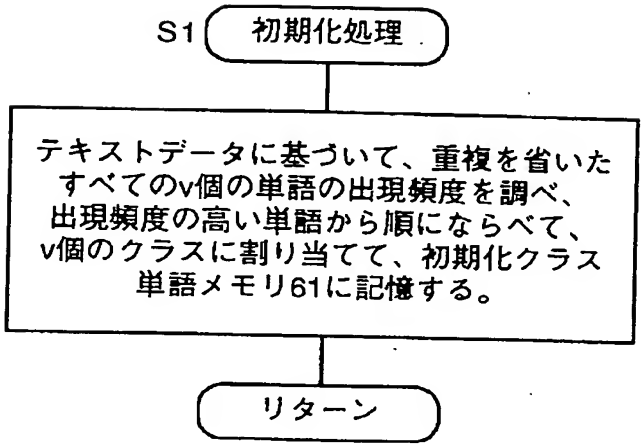
【図８】



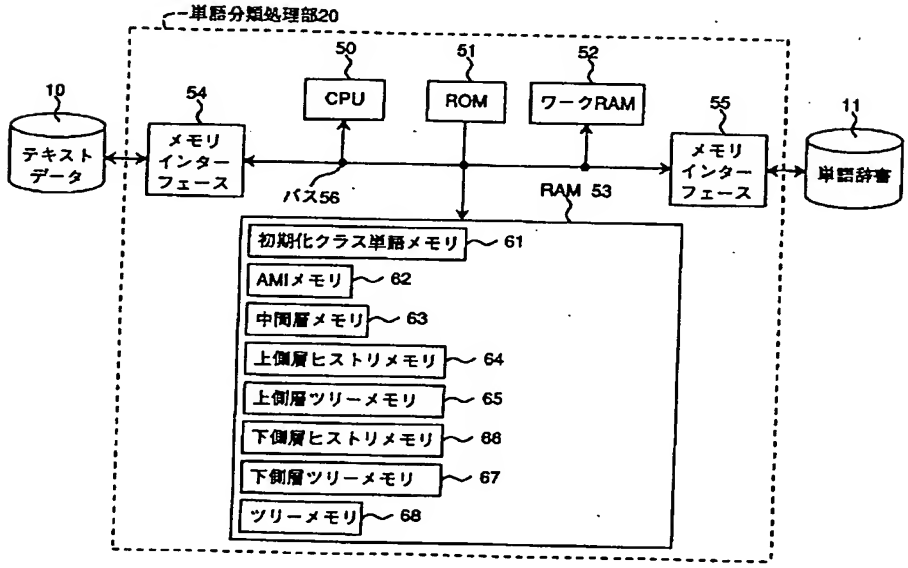
【図９】



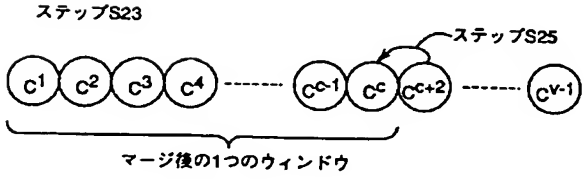
【図１２】



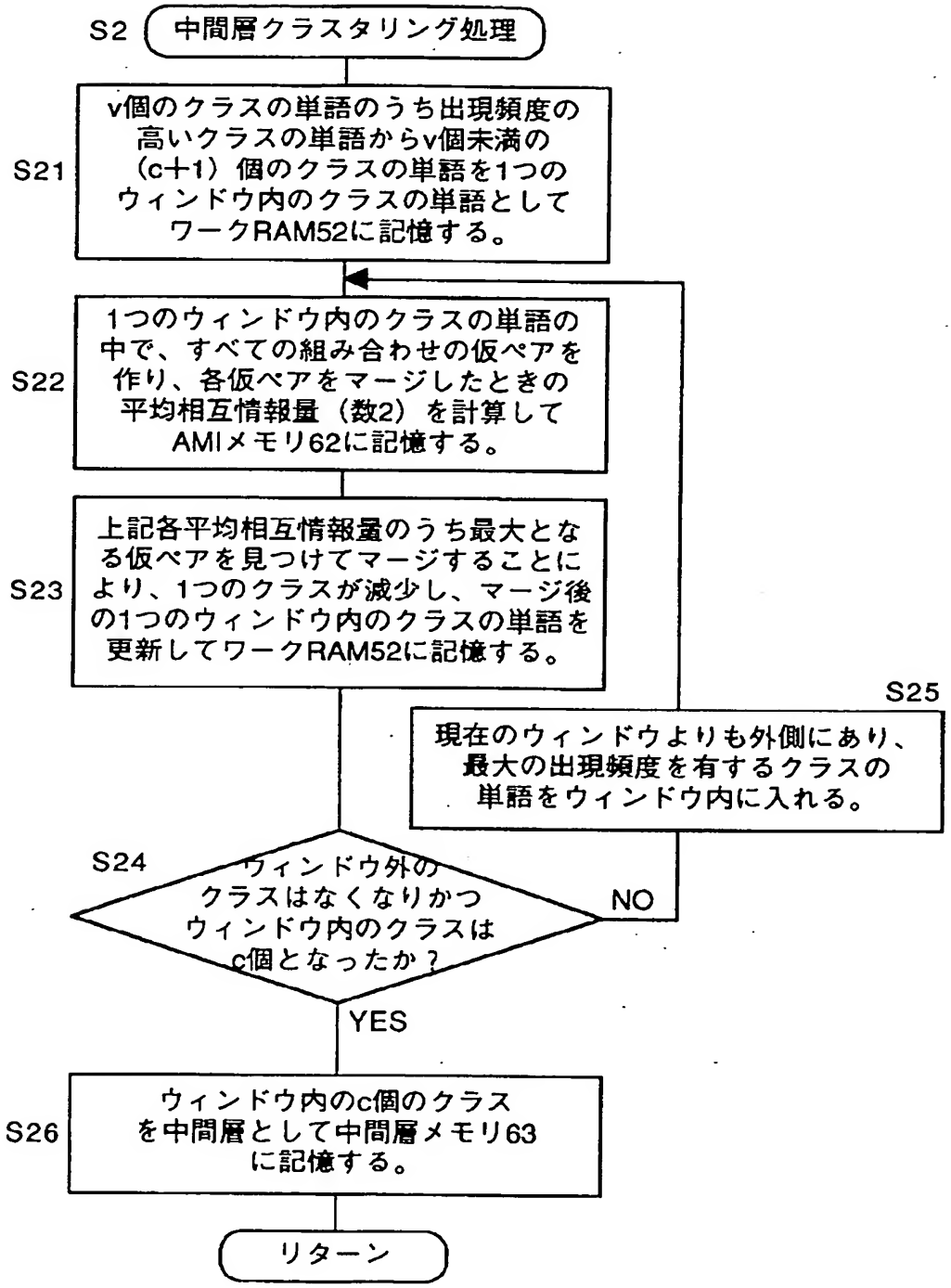
【図１０】



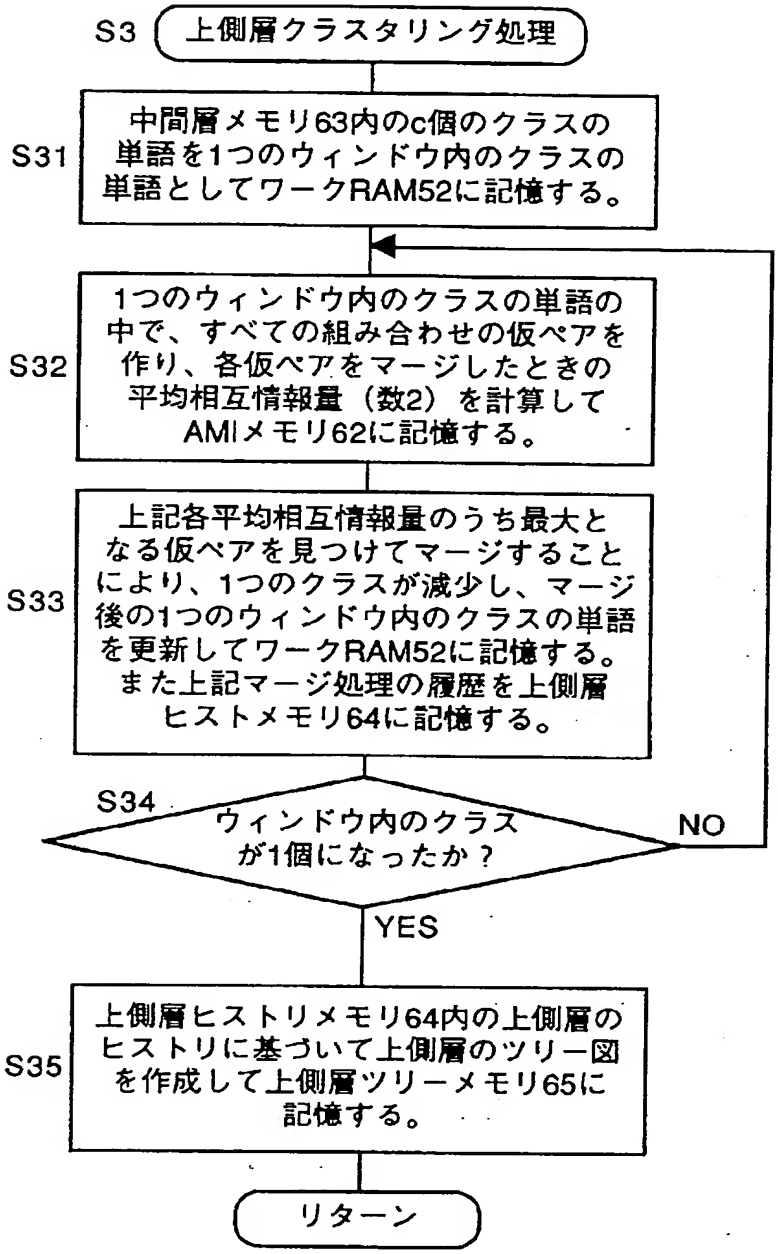
【図１８】



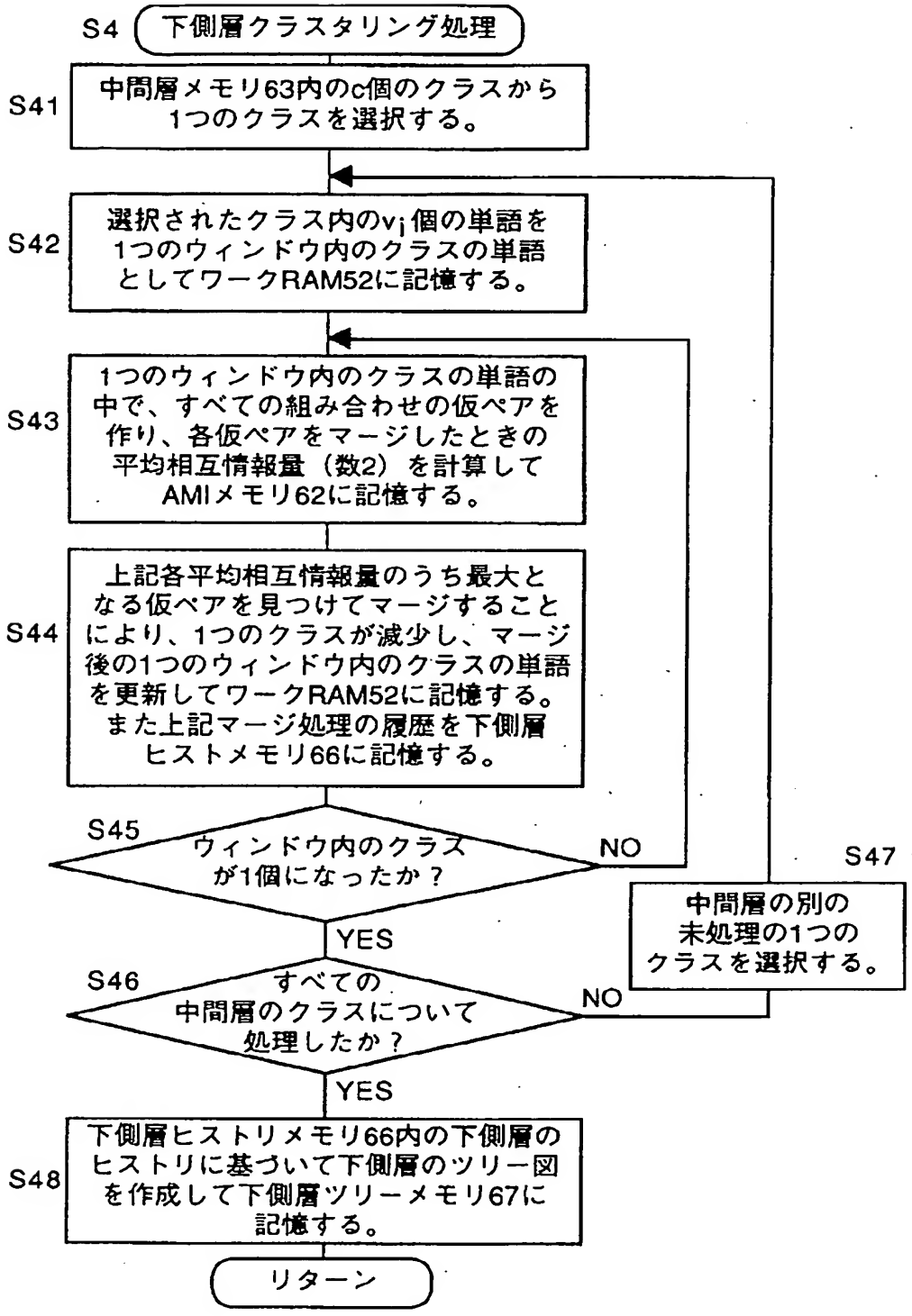
【図１３】


S2
中間層クラスタリング処理
S21
v個のクラスの単語のうち出現頻度の高いクラスの単語からv個未満の（c+1）個のクラスの単語を1つのウィンドウ内のクラスの単語としてワークRAM52に記憶する。
S22
1つのウィンドウ内のクラスの単語の中で、すべての組み合わせの仮ペアを作り、各仮ペアをマージしたときの平均相互情報量（数2）を計算してAMIメモリ62に記憶する。
S23
上記各平均相互情報量のうち最大となる仮ペアを見つけてマージすることにより、1つのクラスが減少し、マージ後の1つのウィンドウ内のクラスの単語を更新してワークRAM52に記憶する。
S25
現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語をウィンドウ内に入れる。
S24
ウィンドウ外のクラスはなくなりかつウィンドウ内のクラスはc個となったか？
NO
YES
S26
ウィンドウ内のc個のクラスを中間層として中間層メモリ63に記憶する。
リターン

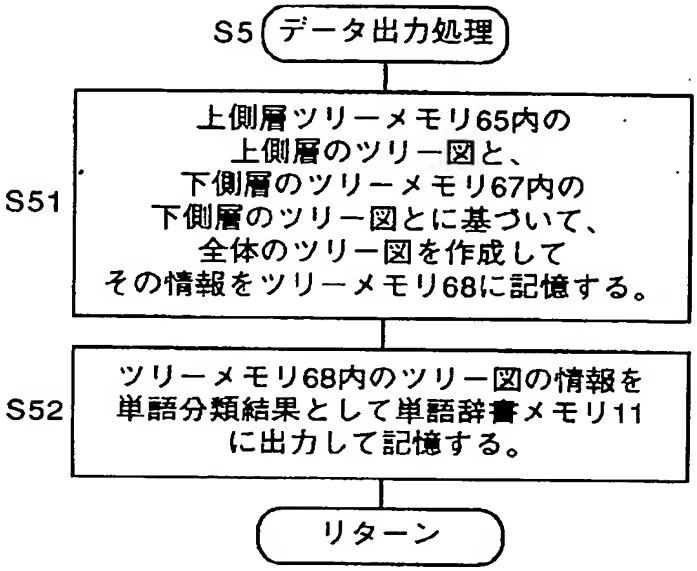
【図１４】


S3
上側層クラスタリング処理
S31
中間層メモリ63内のc個のクラスの単語を1つのウィンドウ内のクラスの単語としてワークRAM52に記憶する。
S32
1つのウィンドウ内のクラスの単語の中で、すべての組み合わせの仮ペアを作り、各仮ペアをマージしたときの平均相互情報量（数2）を計算してAMIメモリ62に記憶する。
S33
上記各平均相互情報量のうち最大となる仮ペアを見つけてマージすることにより、1つのクラスが減少し、マージ後の1つのウィンドウ内のクラスの単語を更新してワークRAM52に記憶する。また上記マージ処理の履歴を上側層ヒストメモリ64に記憶する。
S34
ウィンドウ内のクラスが1個になったか？
NO
YES
S35
上側層ヒストリメモリ64内の上側層のヒストリに基づいて上側層のツリー図を作成して上側層ツリーメモリ65に記憶する。
リターン

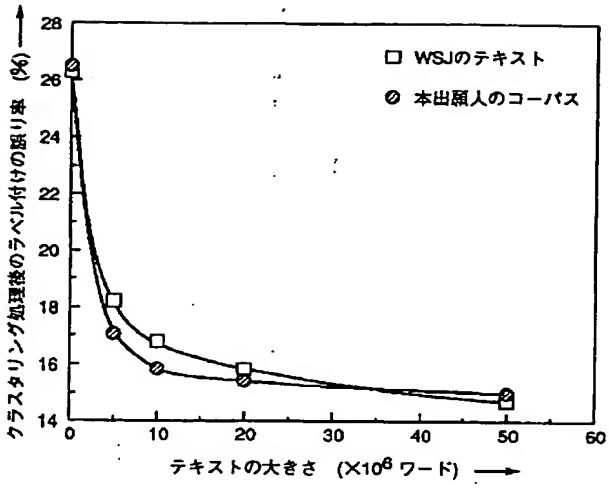
【図１５】


S4
下側層クラスタリング処理
S41
中間層メモリ63内のc個のクラスから1つのクラスを選択する。
S42
選択されたクラス内のvj個の単語を1つのウィンドウ内のクラスの単語としてワークRAM52に記憶する。
S43
1つのウィンドウ内のクラスの単語の中で、すべての組み合わせの仮ペアを作り、各仮ペアをマージしたときの平均相互情報量（数2）を計算してAMIメモリ62に記憶する。
S44
上記各平均相互情報量のうち最大となる仮ペアを見つけてマージすることにより、1つのクラスが減少し、マージ後の1つのウィンドウ内のクラスの単語を更新してワークRAM52に記憶する。また上記マージ処理の履歴を下側層ヒストメモリ66に記憶する。
S45
ウィンドウ内のクラスが1個になったか？
NO
YES
S47
中間層の別の未処理の1つのクラスを選択する。
S46
すべての中間層のクラスについて処理したか？
NO
YES
S48
下側層ヒストリメモリ66内の下側層のヒストリに基づいて下側層のツリー図を作成して下側層ツリーメモリ67に記憶する。
リターン

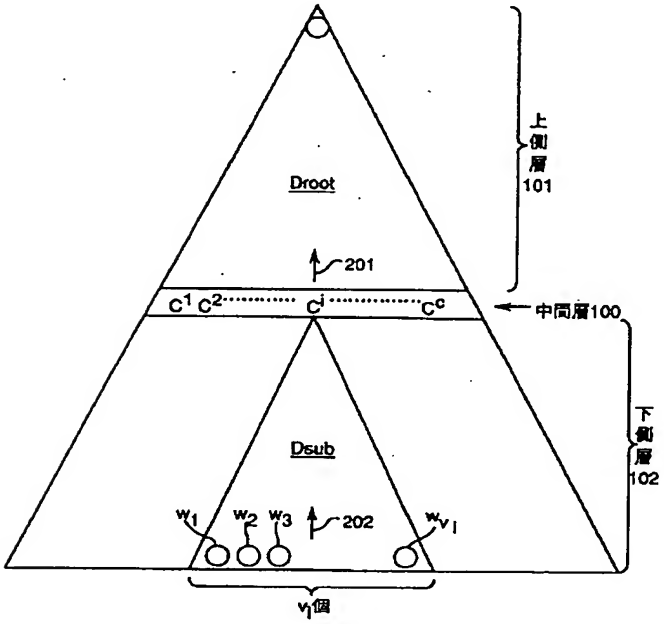
【図１６】


S5
データ出力処理
S51
上側層ツリーメモリ65内の
上側層のツリー図と、
下側層のツリーメモリ67内の
下側層のツリー図とに基づいて、
全体のツリー図を作成して
その情報をツリーメモリ68に記憶する。
S52
ツリーメモリ68内のツリー図の情報を
単語分類結果として単語辞書メモリ11
に出力して記憶する。
リターン

【図２０】


□ WSJのテキスト
⊘ 本出願人のコーパス
クラスタリング処理後のラベル付けの誤り率 (%)
テキストの大きさ (×10⁶ ワード)
28
26
24
22
20
18
16
14
0
10
20
30
40
50
60

【図１９】


Droot
上側層101
201
C¹ C² ····· Cⁱ ····· Cᶜ
中間層100
Dsub
下側層102
w1 w2 w3
202
wv1
v1個